GNN: Projet

1 Démarrer

(a) Téléchargez le graphe des aéroports **augmenté avec les populations des villes**. Vérifiez que vous êtes capables de charger dans un objet pytorch (voir TP précédent pour les lignes a écrire pour que ça fonctionne.)

(b) Vérifier que vous êtes capable d'entrainer sur ce réseau l'un des GCN simples que vous avez écrit lors du premier TP.

2 Consignes du projet

L'objectif du projet est d'écrire un rapport court (Maximum 6 pages) qui décrit la recherche d'une méthode adaptée basée sur les GNN pour répondre à une question de votre choix sur un jeu de données. La première étape consiste à choisir votre question de recherche.

2.1 Questions possibles

- Prédiction de liens: En cachant certains liens, peut-on parvenir à les redécouvrir? Ce problème peut être évalué classiquement en utilisant l'AUC, Average Precision.
- Classification: Vous pouvez avoir comme objectif de prédire l'une des propriétés des nœuds, typiquement le pays ou la population. Ces problèmes peuvent être évalué classiquement en utilisant la precision (classification), la RMSE (regression), etc.
- Détection d'anomalies sur les nœuds/liens: Certaines des informations de pays, et beaucoup d'information de population sont erronnées. La détection d'anomalies se fait en comparant les prédictions du modèle avec les données observées: pouvez-vous découvrir les anomalies ? La validation se fait par observation manuelle, ou en définissant vos propres métriques (ex: pour les 10 villes avec la plus grande différence entre prédiction et observations pour la population, RMSE avec la vrai valeur calculée avec une source externe). La même question se pose pour les liens, certains étant aussi erroné.
- Autre question: si vous pensez à une autre question pertinente, vous pouvez également vous y intéresser

2.2 Attendus

Le projet doit contenir au moins les éléments suivants, classiques pour tout article scientifique de machine learning (et qui devrait l'être dans tous les cas):

- Un tableau qui résume la comparaison entre plusieurs modèles candidats. Ce tableau contient généralement une ligne par méthode, et quelques colonnes correspondant à des scores ou des scénarios différent (ex: colonne pour un test en cachant 10% des liens, une autre en cachant 20% des liens. Pour chacune, une sous-colonne pour l'AUC et pour l'AP pour une tâche de prédiction). En général, les valeurs les plus élevées dans chaque colonne sont indiquées en gras, et si possible une valeur de confiance est donnée, si l'expérience est répété plusieurs fois
- Une "ablation study": pour le modèle qui est votre proposition finale, il vous faut montrer que toutes ses composantes sont utiles. Vous devez donc essayer de remplacer chaque élément "customisé" par un élément plus standard, et montrer que chaque élément customisé est utile. Par exemple, si vous avez utilisé 1 layer GCN et un layer VGAE, montrer qu'enlever le GCN fait baisser les performances. Idem si vous avez proposé une loss customisé, une transformation des données, etc. globalement, montrer que le modèle n'est pas "inutilement complexe".